PSILENCE: A Pseudonymization Tool for International Law

Luis Adrián Cabrera-Diego and **Akshita Gheewala**
Jus Mundi, 30 Rue de Lisbonne, Paris, 75008, France
a.cabrera@jusmundi.com

Abstract

Since the announcement of the GDPR, the pseudonymization of legal documents has become a high-priority task in many legal organizations. This means that for making public a document, it is necessary to redact the identity of certain entities, such as witnesses. In this work, we present the first results obtained by PSILENCE, a pseudonymization tool created for redacting semi-automatically international arbitration documents in English. PSILENCE has been built using a Named Entity Recognition (NER) system, along with a Coreference Resolution system. These systems allow us to find the people that we need to redact in a clustered way, but also to propose the same pseudonym throughout one document. This last aspect makes it easier to read and comprehend a redacted legal document. Different experiments were done on four different datasets, one of which was legal, and the results are promising, reaching a Macro F-score of up to 0.72 on the legal dataset.

1 Introduction

Although the redaction of sensitive information in different types of documents is a common practice in multiple domains, since the announcement of the GDPR and especially after its implementation, the need to find automatic or semi-automatic ways to redact documents has become a priority in many organizations. Historically, the redaction of documents has been done mostly by hand, following guidelines and, in some cases, pattern-matching tools. However, due to its nature, the redaction process is not only slow, but it is also expensive as in many cases an expert needs to be consulted.

In certain domains, like biomedicine, the automatic redaction of documents is well-known thanks to shared tasks, e.g. Stubbs and Uzuner (2015). However, in the legal domain, as in many others, the automatic redaction of documents is still a challenge. For instance, legal documents tend to be long and they have multiple types of entities, e.g. parties, witnesses, experts, judges, lawyers, and citations. Furthermore, some of these entities can be either individuals or organizations. Finally, as we get farther from the beginning of the document, entities become less clear to identify correctly.

Currently, in the legal domain, we can find two different redaction processes: anonymization and pseudonymization, and while both terms are similar, they differ in key aspects. Mourby et al. (2018) summarizes GDPR definition of pseudonymization as the task that "prevents direct identification through attribution, but not through any other mean". Certain organizations add to the definition of pseudonymization the use of a unique identifier for each individual across multiple data sources, that hides their actual identity (Graham, 2012; Elliot et al., 2020). Furthermore, it is a process, that if necessary, can be reversed (Elliot et al., 2020) as only the individuals are substituted, regardless of the occurrence of other elements which could reveal, for example, the gender or age of a person (Allard et al., 2021). In contrast, the goal of anonymization is to remove the complete link between individuals and data (Graham, 2012). Moreover, it is a process that should make the re-identification of people hard to achieve, sometimes by doing additional alterations to the source (Elliot et al., 2020).

We present in this work the first results of *PSILENCE (PSeudonymization of International Law casEs using* NER *and Coreference rEsolution)*, a tool created to pseudonymize international arbitration documents in English. These first results come from multiple experiments done over four different datasets, one of which has been created by a group of legal experts for this specific task.

The rest of the paper is organized as follows. We present the scope and objectives of this work in Section 2. In Section 3, we present the most relevant works found in the literature related to the automatic redaction of documents, i.e. methods

Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization
CALD-pseudo 2024, pages 25–36
March 21, 2024 ©2024 Association for Computational Linguistics

and data, as well as some additional relevant tasks. Then, we present the methodology of our system in Section 4. The data collection explored in this work is described in Section 5 while the evaluation setup is detailed in Section 6. The experimental results and their discussion are presented in Section 7 and Section 8 respectively. Finally, we conclude and propose our future work in Section 9.

2 Scope and Objectives

PSILENCE has been developed to semi-automatize the pseudonymization process of English documents within Jus Mundi[1]. Currently, it focuses only on entities of type people, however, we are aware of the existence of other types of information that need to be hidden, such as emails and addresses. Furthermore, from all the entities of type people, only those of type witnesses are redacted. This means that we do not redact lawyers, judges, or parties.[2]

Therefore, PSILENCE has two main goals. First, to propose to a legal expert a list of people that should be redacted in the document to keep the sensitive information hidden. Secondly, to cluster the names of people to provide a unique identifier to each redacted person within a document. This means that different name variations of the same person are grouped together. For instance, *"Mariano Puerta"* and *"Mr. Puerta"* will compose one cluster, while *"Laura Puerta"* would be put into a different one. In this way, we can simplify the redaction process and improve the readability and comprehension of a redacted document.

3 Related Work

In the health and biomedical domains, we can find multiple tools developed for the anonymization, pseudonymization, and deidentification of information, as presented by Chevrier et al. (2019) and Leevy et al. (2020). However, in the legal domain, there is a reduced number of works. For instance, we can name ANOPPI (Oksanen et al., 2019; Arttu Oksanen et al., 2022), a pseudonymization tool for Finnish Court documents that makes use of multiple NER systems, based on rules and machine learning. It uses regular expressions and dictionaries to find elements such as registration plates or specific names. As Finnish inflects pronouns and nouns, they perform morphological analysis to correctly inflect pseudonyms. Individuals are not grouped, this means that each occurrence of them is assigned a different identifier. In Schamberger (2021), the authors present an anonymization tool for German court rulings. Specifically, the authors create a NER system by using BERT embeddings (Devlin et al., 2019) through a BiLSTM and CRF architecture. Pilán et al. (2022) compare different tools for anonymyzing legal documents: Presidio[3], a generic NER system based on RoBERTa (Liu et al., 2019) and a specialized NER based on Longformer (Beltagy et al., 2020).

Outside the legal domain, we can highlight the work of Biesner et al. (2022). In this paper, the authors present a full anonymization system for German financial documents. The system considers the anonymization task as a sequence tagging problem, thus, they make use of NER for detecting entities. They explored elements such as word embeddings, contextual embeddings, and different neural network architectures for creating the NER system. Similarly, Papadopoulou et al. (2022) use knowledge graphs and k-anonymity (Sweeney, 2002) to generate a weakly supervised dataset. Then, the generated dataset is used to fine-tune RoBERTa (Liu et al., 2019) and create an anonymization tool following a NER architecture.

Regarding pseudonymization and anonymization data there are not many publicly available datasets. The documents that need this kind of tool have to be pseudonymized or anonymized before becoming public, due to privacy reasons, and, annotating documents is an expensive and time-consuming task. Thus many of the legal datasets used in the literature are private, such as the works of Barriere and Fouret (2019) and Garat and Wonsever (2022). One exception is the TAB Corpus (Pilán et al., 2022), which is a collection of publicly available documents from the European Court of Human Rights that have been annotated for evaluating anonymization tasks.

Outside the legal domain, there are clinical datasets such as the *2014 i2b2/UTHealth* corpus (Stubbs and Uzuner, 2015) and the *2016 CEGS N-GRID Shared Task* (Stubbs et al., 2017). Papadopoulou et al. (2022) created an anonymization dataset using a collection of Wikipedia biographies.

As the pseudonymization and anonymization

[1] https://jusmundi.com/

[2] This was defined by Jus Mundi's legal team according to their needs. However, PSILENCE is capable of redacting all types of person entities if necessary.

[3] https://github.com/microsoft/presidio

tasks can be seen as a NER one (Pilán et al., 2022; Garat and Wonsever, 2022; Papadopoulou et al., 2022) it is not uncommon for researchers to use general NER datasets, such as CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003), and then apply the (pre-) trained models through a zero-shot approach into the legal domain. This is the case of the works presented by Schamberger (2021) and Pilán et al. (2022). However, as Pilán et al. (2022) conclude, the zero-shot results are not the best as, in some cases, the entities to mask are different than those available in the original tagset.

Although the clustering of individuals for the pseudonymization and anonymization tasks has been considered relevant in some works (Pilán et al., 2022; Garat and Wonsever, 2022), the amount of available resources regarding this aspect is scarce. For instance, in TAB (Pilán et al., 2022) only 1.7k of 24k entities of type person belong to a cluster, the rest are singletons[4]. In the case of CoNLL 2012 coreference corpus (Pradhan et al., 2012) there are no singletons. The best exception is *LitBank* (Bamman et al., 2020), a collection of 100 fiction documents in English that are annotated with coreference resolution, and presents singletons and clusters.

Finally, we can find some additional tools in the literature related to the pseudonymization task. In Gupta et al. (2018), the authors present a tool for identifying parties of legal cases using NER and coreference resolution. Moreover, in Kalamkar et al. (2022), the authors present a NER system for annotating Indian legal documents on which they reconcile types of named entities using rules and coreference resolution. BookNLP[5] a Spacy-based tool created for processing long documents, especially fiction books. Among BookNLP's tools, we can name a character clustering and a coreference resolution module. Finally, PeTra (Toshniwal et al., 2020) is a model based on BERT (Devlin et al., 2019) which uses memory modules to keep track of people within short documents.

4 Methodology

In Figure 1, we present PSILENCE's architecture, which is composed of four modules. In the first module, we make use of a Python-based HTML parser and Spacy (Honnibal et al., 2020) to pre-process the documents. During the pre-processing of documents, we convert HTML documents into plain text divided into paragraphs and we extract citations that were defined as HTML spans. The second module is a hybrid NER system, i.e. based on a machine learning model and rules; its goal is to detect different types of named entities within a document. The third module is a simplified coreference resolution model that only clusters names of entities and does not consider any kind of pronouns.[6] The fourth module is a reconciliation system, similar to the one used in Kalamkar et al. (2022), which tries to determine the exact type of entity in a document, even if the context in which it occurs, is not clear.

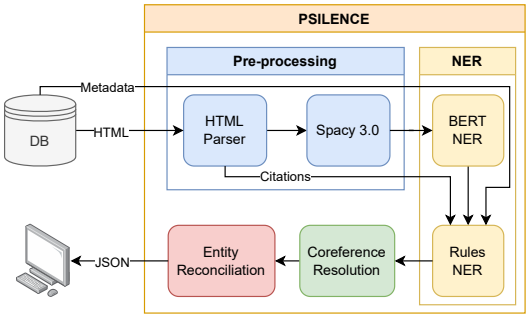

Figure 1: Global architecture of PSILENCE.

At the end of the pipeline, we create a pseudonymization dictionary, which is a JSON file, see Figure 2, indicating the different clusters found for each type of person entity. Each cluster contains all the variations found for the same person with their occurrences based on character positions. Based on the example presented in Figure 2, the occurrences of the names *"Bill Scott"*, *"William Scott"* and *"Scott"*, would be replaced with *"WITNESS_1"* while the name *"McConnell"* would be replaced with *"WITNESS_2"* in the pseudonymized document. Although in this work we focus on clusters of type Witness, we provide other types of clusters in the JSON output in case we make a mistake in the grouping or classification of entities.

The second, third, and fourth modules will be described in detail in the following subsections.

4.1 Named Entity Recognition (NER)

For extracting named entities, PSILENCE uses a hybrid NER system. It was done by coupling a

[4]A singleton is a type of cluster composed of only one person occurring only once in a document.

[5]https://github.com/booknlp/booknlp

[6]The reasons for not considering pronouns is that it makes the coreference resolution task harder to do and, as Pilán et al. (2022) indicate, pronouns do not tend to leak highly sensitive information even in anonymization tasks.

```
{
 "clusters": {
  "WITNESS": [
   {
    "Bill Scott": [[2003,2013]],
    "William Scott": [[2317,2330]],
    "Scott": [[2443,2448], [3305,3310]]
   },
   {
    "McConnell": [[3300,3309]]
   }
  ],
  "LAWYER":[
   {
    "Bermudez": [[1712,1720]]
   }
  ]
 }
}
```


Figure 2: Example of a PSILENCE's JSON output file. The file presents the different person entity types and the clusters found in the document. We indicate as well the character position in which the replacement needs to be done.

machine learning model, through a zero-shot approach, and a collection of rules.

The machine learning model is a transformer-based NER system trained on CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) using $\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2019).[7] This model was used due to the lack training data[8] and it can predict four types of named entities, person, organization, miscellaneous, and location.

Regarding the collection of rules, we use regular expressions and string matching[9] to determine whether an entity found by the machine learning approach should be specialized. For instance, we do string matching between named entities of type person and metadata from the case database to find the names of judges, lawyers, and parties. In the case of authors, we use, for example, regular expressions to extract them from citations found in the pre-processing module.

In total, we can detect 12 types of entities: Party, Judge, Lawyer, Arbitrator, Tribunal member, Expert, Author, Law firm, Person, Organization, Miscellaneous, and Location. These are obtained using the following approaches. *Machine learning* - Person, Organization, Miscellaneous, and Location. *Metadata string matching* - Party, Judge, Lawyer, Arbitrator, Tribunal member, Expert, and Law firm. *Regular expressions* - Judge, Lawyers, Authors. As it can be seen, the category Witness is not present in the aforementioned list, this is because all the entities of type Person that could not become specialized, e.g. Judge and Author, are considered as witnesses at the end of PSILENCE's pipeline.

Although the approach described before can be counter-intuitive, i.e. to find specialized named entities rather than directly finding witnesses, it should be indicated that finding only witnesses is harder. In the first place, and especially as we get farther from the beginning of the document, the context in which the name of a person occurs might not be descriptive enough to determine whether it is a witness or not. At the beginning of a document, specialized people tend to be formally introduced either using titles or specific contexts. For instance, a lawyer can be introduced in a document as *"Doe QC"*, a judge as *"Honorable Doe"*, or an arbitrator as *"Arbitrator: Ms. Jane Doe"*. However, in the case of witnesses, these do not tend to be introduced directly as witnesses, such as in *"Mr. Doe was a personal trainer in the defendant's company and noticed that..."*. In the second place, this approach makes PSILENCE's output easier to correct by humans, for example, if a person is wrongly marked as an Expert, all their occurrences in a document can be easily converted into Witness, without having to find these by hand. Finally, it makes PSILENCE easier to use in different legal contexts, such as those where judges or lawyers need to be redacted as well.

4.2 Coreference resolution

For clustering named entities, we use a coreference resolution system based on the work of Clark and Manning (2016a,b). This means that it is composed of a mention-pair encoder, a cluster-pair encoder, a mention ranking model, and a cluster ranking model; moreover, the neural network has three fully connected hidden ReLU layers.

The input features of the neural network are presented as follows:

- Dense Embeddings: We use FastText with subword information (Bojanowski et al., 2017) to vectorize entities and entities contexts. Specifically, we use those trained on Common Crawl[10] and we reduced the size of

[7] This model was not trained by us, instead it was downloaded from https://huggingface.co/dbmdz/bert-large-cased-finetuned-conll03-english

[8] This research was done before the publication of Kalamkar et al. (2022), which proposes an English Indian Legal NER dataset. Moreover, it should be noted that we focus on international arbitration cases and not national legal cases as it happens in Kalamkar et al. (2022).

[9] This is done using RapidFuzz: https://github.com/maxbachmann/RapidFuzz.

[10] crawl-300d-2M-subword.zip

the embedding from a dimension of 300 to 100 using FastText API[11].

- Length of named entity: Using binary encoding, we set the number of characters in a named entity.
- Named entity location: It is the relative position of the named entity within a document.
- Matching root: Using Spacy's dependency parser, we compare whether the root of a named entity matches the root of other ones.
- Words intersection: Proportional number of words shared between couples of named entities.
- Exact match: We compare whether two named entities have an exact match.
- Relaxed match: We make use of RapidFuzz to determine the degree of string matching between a couple of named entities. In other words, we utilize fuzzy string matching metrics as digitized legal documents can contain misspelling mistakes, originated either by the OCR or by the data entry clerk.
- Cosine similarity: Using FastText embeddings, we calculate the cosine similarity between named entities.
- Named entity distance: We calculate the relative distance, in terms of words, between a couple of named entities.
- Dense representation of context: We calculate, using FastText embeddings, the dense representation of the named entities' contexts.

All the string comparisons are done using UTF-8 and ASCII encodings to prevent mistakes by the use of diacritics.

In Table 1, we present the hyperparameters used for training the coreference resolution model.

It should be indicated that during prediction time, we pre-cluster the named entities using RapidFuzz and a similarity score of 0.6. This allows us to decrease the processing time on long documents. This approach is similar to the one used by BookNLP[12] for the coreference resolution (Bamman et al., 2020).

4.3 Entities reconciliation

One common problem in NER tasks, especially in long documents, is the fact that certain entity names can be predicted with different types in multiple paragraphs or sentences (Kalamkar et al., 2022). The main reason is that the context in which an entity occurs might change. For instance, at the beginning of a document, it might be stated that *Mr. X* is a lawyer but, later on in the document it is just presented as *Mr. X*. In these cases, it might be impossible to determine the correct entity type, not only for humans (without reading the full document), but for machine learning models too. Therefore, it is necessary to reconcile entity types to have the best performance possible.

Table 1: Hyperparameters used for training the coreference resolution model.

Hyperparameter	**Value**
Maximum Epochs	200
Early Stop Patience	30
Learning Rate	0.001
Scheduler	Linear with warm-up
Warm-up Ratio	0.1
Optimizer	AdamW with bias correction
AdamW ϵ	1×10^{-8}
Random Seed	1111
Dropout rate	0.5
Weight decay	0.01
Embeddings size	100
h1 size	1000
h2 size	500
h3 size	500
Cost False New	0.8
Cost False Anaphoric	0.4
Cost Wrong Link	1.0

In this work, we use the output generated by the coreference resolution system along with some rules to reconcile entities. Specifically, for a given cluster of people, we start by counting the different types of entities. If only one type of entity exists, we consider the type of entity to be correct. However, if it is the opposite, i.e. more than one type, we use the following rules:

- If one of the entities is marked as a party, then all the entities become of type party and will be ignored for the pseudonymization process.
- If more than 30% of the entities are not of type person, i.e. Location, Miscellaneous, Law Firm, or Organization, the cluster will be ignored for the pseudonymization process.
- If the most frequent type of entity is a Judge, Author, Expert, Arbitrator, Tribunal Member, or Lawyer, then the cluster is ignored for the pseudonymization process.

The clusters considered to pseudonymize, i.e. Witness, are those of type Person that after the reconciliation process could not be specialized. These rules were developed and fine-tuned experimentally by

[11] https://fasttext.cc

[12] https://github.com/booknlp/booknlp

Table 2: Statistics of the legal corpus.

Per document	Median	Minimum	Maximum
Tokens	10 809	496	112 509
Witness entities	14	1	305
Clusters	4	1	38

assessing the performance of PSILENCE on the development part of a legal corpus (see Section 5).

5 Data

For training PSILENCE coreference resolution system, we use three different corpora. *LitBank* (Bamman et al., 2020) is the main training corpus because it contains singletons, documents are long and it is one of the biggest coreference resolution corpora. However, due to its literary nature, we decided as well to use two entity-linking-related corpora, *In Media Res* (Brasoveanu et al., 2020) and *AIDA-CoNLL-Yago* dataset (Hoffart et al., 2011); both of these corpora focus on news articles. Even though these two corpora are not annotated with coreference resolution groups, as we focus only on the clustering of people, we make use of their entity-linking annotations to determine clusters. In other words, named entities of type person can be grouped thanks to common knowledge-base links. For example, in AIDA-CoNLL-Yago, in a document talking about the signer *"Johnny Allen Hendrix"*, all his name variations, e.g. *"Hendrix"* and *"Jimi Hendrix"*, are linked to the same knowledge base Yago ID. To improve the quality of these two last corpora, we manually validated some of the clusters, and in the case of AIDA-CoNLL-Yago, we also included some heuristics to match some people that were not linked correctly.[13] For the three corpora, we use a training, development, and testing partition; therefore, we can fine-tune the models and evaluate their performance.

Besides, we have a collection of 140 international arbitration documents written in English covering different types of cases: sports (121), commercial (8), inter-state (5), Iran-US claims (2), and investor-state (4); see Table 2 for statistics. These 140 documents were manually annotated by a group of expert lawyers at Jus Mundi. Specifically, these experts created for each document a list of witness clusters; in other words, they found all the witnesses in a document and grouped their different occurrences into clusters.[14] It should be indicated that these documents are in HTML format and were previously enriched with citations using an in-house tool. Each document is associated with metadata which was manually verified by Jus Mundi's legal team. From the 140 documents, 33 were used for fine-tuning PSILENCE's pipeline, i.e. NER, and coreference resolution, and 107 were used for testing it.

6 Evaluation

In this paper, we use the evaluation framework proposed in CoNLL 2012 Coreference Shared Task (Pradhan et al., 2012). It asses in the first place whether all the named entities have been found within a document. And, in the second place, it evaluates how well these entities have been grouped into clusters. This evaluation framework is composed of three metrics, B^3 (Bagga and Baldwin, 1998), CEAF (Luo, 2005) and MUC (Vilain et al., 1995). However, instead of using MUC as defined by Vilain et al. (1995), we make use of a modified version that takes singletons into account. Specifically, for singletons, we define the minimum number of correct links as $|k(S)|$, instead of $|k(S)| - 1$, and the number of missing links as $|p(S)|$ instead of $|p(S)| - 1$; where $|k|(S)$ is the size of the key cluster for mention S, i.e. 1, and $p(S)$ is the intersection of the predicted cluster and the key cluster of mention S. In simple words, MUC for singletons becomes a binary metric. This change was necessary as the CoNLL 2012 Coreference Shared Task did not consider singletons but our four corpora do.

As indicated in Section 3, there are not many available tools for pseudonymizing legal documents. However, we compare PSILENCE coreference resolution tool with BookNLP[15]. Specifically, we assess the clustering performance of PSILENCE and BookNLP when they are provided with the gold standard entities, i.e., those that have to be pseudonymized. We use BookNLP because it was designed to process long documents and it is capable of performing coreference resolution (Bamman et al., 2020).[16] The comparison is done

[13] In AIDA-CoNLL-Yago, we used the original documents. However, for In Media Res, we create pseudo-documents by grouping sentences based on the co-occurrence of people.

[14] We did not include clusters of other types of entities in these lists, such as lawyers or judges, as they were out of the project scope. But also because their annotation would have become harder to achieve.

[15] https://github.com/booknlp/booknlp

[16] Although BookNLP has its own NER, we did not adapt its NER to predict and/or filter subtypes of people, like lawyers and judges, due to the complexity of the task.

on the legal documents, thus, we can assess how well the tools behave in the legal domain when all the correct entities are given.

7 Results

We present in Table 3 the results, in terms of Macro F-score, obtained by our coreference resolution system, when it was applied on the testing partitions of AIDA-CoNLL-Yago, In Media Res and LitBank. The results presented in Table 3 show us how well can we cluster the names of people in different types of documents and circumstances. It is clear that as the length of the documents increases, as it happens in LitBank, the performance decreases.

In Table 4 and Table 5, we show the F-scores obtained by our coreference tool and BookNLP, the baseline, regarding the clustering of gold standard entities, i.e. the names of people that had to be pseudonymized, over the legal development and testing corpora respectively. The macro outcomes presented in both Table 4 and Table 5 show that, despite applying the coreference resolution tools to an unseen domain, they manage to cluster people correctly in most documents. Nonetheless, the micro outcomes shown in Table 4 and Table 5, indicate that in documents where a great number of people co-occur, the performance decreases as it is harder to disambiguate people.

In Table 6, we introduce PSILENCE pipeline's results. We can observe in Table 6 that when we include the NER system into the pipeline, the performance of our coreference resolution tool is affected. This means that the detection of named entities is not perfect and the produced noise affects the clustering of people. Specifically, in the test corpus we pass from a macro CoNLL F-score of 0.95 (Table 5) to 0.82 (Table 6).

8 Discussion

Regarding the results presented in Table 3, we can observe that the macro F-scores achieved by the coreference system tend to be greater than 0.90, meaning that in general, most of the documents are clustered correctly. The performance decreases as the length of the document increases because the number of mentions increases, thus the number of pairs needed to be compared increases as well. Moreover, the documents from AIDA-CoNLL-Yago and In Media Res are relatively small and have fewer named entities and clusters than LitBank.

As we observed in Table 4 and Table 5, our coreference resolution system, performed in general, better than the one found in BookNLP. This can be due to several aspects. In the first place, PSILENCE coreference resolution system was trained on two more datasets. This means that PSILENCE was trained on more examples but also from different domains, literary and news. Secondly, to use BookNLP as a baseline, we had to introduce our gold standard named entities into BookNLP, meaning that we had to remove their NER system and modify certain pipelines. This could have affected the performance; also, BookNLP was designed to link personal pronouns to names too. Moreover, we do not see any change between BookNLP's small and big models (Table 4 and Table 5).

We performed a manual analysis of certain clusters found in our legal dataset to better understand Table 4 and Table 5. From this analysis, we determined that there are recurrent errors that occur in both PSILENCE and BookNLP. We found out that spelling name variations are one of the most common reasons for people not being correctly clustered. For instance, *"Mahmood"* can also be referred to as *"Mahmoud"*; *"Lief"* as *"Liefs"* and *"Kuan"* as *"Koan"*. Another frequent clustering error across both approaches occurs when the full name is used but then, only a part of it is used later in the text, like *"Michael S. Blatter"* as *"Blatter"* and *"Lalit Merchant"* as *"L Merchant"*. We noticed a drop in performance when the documents have people with long names, such as double last names, but also if they contain accentuated letters. Nonetheless, we also noticed that with PSILENCE, we can correctly cluster some entities among the above-mentioned instances. For instance, we manage to cluster *"Bill Essick"* and *"William Essick"* correctly whereas they remain as separate entities with BookNLP. It should be indicated that these types of errors are not uncommon, neither in PSILENCE or BookNLP. We believe it is related to the sentences' context, but a deeper analysis is needed.

Some of the previous errors might be able to be fixed by changing the embeddings type, from word to contextual ones like those provided by BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019). However, this would mean that the architecture of the coreference resolution system would need to change completely, as contextual embeddings are not designed for single-word analysis, and have to be trained differently for calculating cosine similarity (Reimers and Gurevych, 2019). Moreover, mod-

Table 3: Results in terms of macro F-score for each testing partition of the corpora used for training the coreference resolution system.

Corpus	MUC	BCUB	CEAFE	CoNLL
AIDA-CoNLL-Yago	0.98	0.97	0.96	0.97
In Media Res	0.94	0.95	0.92	0.94
LitBank	0.96	0.93	0.80	0.90

Table 4: Results of the coreference resolution task in terms of F-score, micro and macro averaged, for legal development corpus.

System	MUC		BCUB		CEAFE		CoNLL	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Ours	**0.90**	**0.95**	0.50	**0.95**	**0.64**	**0.90**	0.68	**0.93**
BookNLP Small	0.89	0.94	**0.61**	0.92	0.58	0.86	**0.69**	0.91
BookNLP Big	0.89	0.94	**0.61**	0.92	0.58	0.86	**0.69**	0.91

els such as Sentence BERT (Reimers and Gurevych, 2019) have been created to find similar sentences and not similar words, unlike FastText.

As we observed in Table 6, the coreference resolution F1-score decreases by 37% (micro) and 24% (macro) in comparison to the clustering-only task. This means that PSILENCE's NER has trouble in correctly detecting all the different types of named entities. For instance, the machine learning model sometimes cannot find all the entities in a sentence, or if they are found they can be tagged with the wrong type, or they are split into multiple smaller ones, or the boundaries are wrong, e.g. *"Romano F."* and *"Subiotto Q.C."* rather than *"Romano F. Subiotto"*. Regarding the collection of rules, sometimes it is hard to correctly apply them. For instance, in some documents, the parties were stated as *"Company (Country) Ltd."* or lawyers as *"John R. Doe"*, however in our metadata, these entities were *"Company Ltd."* and *"John Roe Doe"*.

To solve the aforementioned issues, we can propose certain solutions. First, we need to reduce our dependency on rules for the NER by training a specialized legal NER rather than using a generic one in a zero-shot way. Secondly, to reduce the number of entities with wrong boundaries, the new NER should be trained with a CRF layer, like in Ma and Hovy (2016), and use an IOBES encoding, as in Ratinov and Roth (2009). Also, we might need to use data augmentation methods, such as in Cabrera-Diego and Gheewala (2023), where a frustratingly easy domain adaption method is used to mix different legal NER corpora.

Moreover, some of the detected errors were caused by the reconciliation module. In other words, the rules used in this module were not robust enough to detect or solve issues generated by the NER model. For example, in one document a law firm was incorrectly tagged as a person rather than as an organization; in this case, the reconciliation module determined that the entity was of type person because it was the most frequent type, thus it was an entity that had to be pseudonymized.

Some other errors found during the analysis were caused by a wrong splitting of sentences. This was particularly noticeable when a paragraph contained citations that were not tagged in the HTML document, which in consequence made a paragraph be split into wrong sentences. In consequence, authors found in these undetected and wrongly split citations were considered many times as witnesses because specialization rules could not be applied. Other splitting errors in sentences come from the fact that Spacy, was not trained to analyze legal documents, thus it is not aware of specialized abbreviations such as *Hon'ble* and *Q.C.* Moreover, we found out that in general, Spacy is bad at processing long sentences, such as those that are found in legal documents. Therefore, when a paragraph is wrongly split into sentences, it has a consequence not only on the NER system but also on the coreference resolution one. To solve these errors, one option is to train our model for splitting sentences, although it can be complicated to achieve due to the number of data necessary to train this kind of model. Another option is to stop using sentences

Table 5: Results of the coreference resolution task in terms of F-score, micro and macro averaged, for the legal testing corpus.

System	MUC		BCUB		CEAFE		CoNLL	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Ours	**0.94**	**0.97**	0.47	**0.96**	**0.45**	**0.92**	**0.62**	**0.95**
BookNLP Small	0.9	0.93	**0.51**	0.92	0.37	0.85	0.59	0.90
BookNLP Big	0.9	0.93	**0.51**	0.92	0.37	0.85	0.59	0.90

Table 6: Results of the pseudonymization pipeline, i.e., NER, coreference resolution, and entities reconciliation, in terms of F-score, micro and macro averaged, for the development and testing corpora.

Corpus	MUC		BCUB		CEAFE		CoNLL	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Developement	0.70	0.87	0.24	0.86	0.32	0.80	0.42	0.82
Test	0.71	0.77	0.18	0.77	0.27	0.71	0.39	0.72

for delimiting the context in which a named entity occurs. However, this would mean that it could be harder to determine the actual context of a named entity in the coreference resolution system.

Despite the complexity of the pseudonymization task and the use of multiple deep learning models through a zero-shot approach, we consider the macro results shown in Table 6 good in general. Nonetheless, there is still work to be done, especially when we observe the micro results (Table 6). These results indicate that we need to continue working on the clustering of people in long documents because it becomes harder to keep track of people. We might need to explore more complex methods for clustering people using memory systems, such as PeTra (Toshniwal et al., 2020). However, we also need to consider that many of the works of coreference resolution are done on relatively short documents.

9 Conclusions

In this paper, we presented the first results of PSILENCE, a pseudonymization tool for the semi-automatic redaction of international arbitration documents in English, where people are clustered, to accelerate the human validation step and improve the readability of the document.

Experiments were done on different datasets, including one composed of legal documents. The obtained results were promising, especially for the clustering of people through coreference resolution. For instance, we got a macro F-score of 0.95, when clustering gold standard named entities, and a macro F-score of 0.72 when we use the NER.

An analysis of the results showed that some of the errors come from the fact that we use multiple rules at different levels. But also, because the current implementation of PSILENCE is based on multiple zero-shot approaches, meaning that the training data did not come from the legal domain. Therefore, to improve PSILENCE, it will be necessary to work on a specialized legal corpora.

In the future, we will work on the improvement of the PSILENCE system as discussed in Section 8. Moreover, we would like to cluster named entities through multiple documents to assign them the same pseudonym. This would be useful when a case has multiple documents and certain people occur in several of them, allowing us to increase the readability of complex cases.

Finally, we will train PSILENCE using multilingual language models on legal documents in other languages than English, especially those from the European Union where legal documents are subject to GDPR rules.

Acknowledgments

This work was possible thanks to the granted access of IDRIS (Institut du Développement et des Ressources en Informatique Scientifique) High-performance computing resources under the allocation 2022-AD011012667R1 and 2023-AD011012667R2 made by GENCI (Grand Équipement National de Calcul Intensif).

Limitations

PSILENCE has different limitations that need to be clarified. In the first place, while we indicate that PSILENCE can pseudonymize, if necessary, different types of person entities besides witnesses, it should be stated that we have not evaluated yet how well PSILENCE can detect these other person entities. The main reason is that we do not have those annotations and are very expensive to manually get. While we expect PSILENCE's coreference resolution system to perform similarly to the results presented in this work, we cannot ensure that the quality of the NER will be equal for all the types of named entities. Nevertheless, we expect that by deploying PSILENCE in Jus Mundi, we will be able to have more and better annotations that could be used to train specialized tools. In the second place, we have explored different types of international arbitration cases, however, there are many more. Thus, we cannot ensure that the current pipeline used in PSILENCE can be applied to all types of arbitration, at least without a fine-tuning process.

References

Tristan Allard, Louis Béziaud, and Sébastien Gambs. 2021. Publication of Court Records: Circumventing the Privacy-Transparency Trade-Off. In *AI Approaches to the Complexity of Legal Systems XI-XII*, pages 298–312, Cham. Springer International Publishing.

Arttu Oksanen, Minna Tamper, Jouni Tuominen, Aki Hietanen, and Eero Hyvönen. 2022. A Tool for Pseudonymization of Textual Documents for Digital Humanities Research and Publication. In *6th Digital Humanities in Nordic and Baltic Countries Conference (Book of Abtsracts)*, pages 107–108, Uppsala, Sweden.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation: Workshop on Linguistics Coreference*, volume 1, pages 563–566, Granada, Spain. European Language Resources Association.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An Annotated Dataset of Coreference in English Literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Valentin Barriere and Amaury Fouret. 2019. May I Check Again? — A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 327–332, Turku, Finland. Linköping University Electronic Press.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer.

David Biesner, Rajkumar Ramamurthy, Robin Stenzel, Max Lübbering, Lars Hillebrand, Anna Ladi, Maren Pielka, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. 2022. Anonymization of German financial documents using neural network-based language models with contextual word representations. *International Journal of Data Science and Analytics*, 13(2):151–161.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Adrian M.P. Brasoveanu, Albert Weichselbraun, and Lyndon Nixon. 2020. In Media Res: A Corpus for Evaluating Named Entity Linking with Creative Works. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 355–364, Online. Association for Computational Linguistics.

Luis Adrián Cabrera-Diego and Akshita Gheewala. 2023. Jus mundi at SemEval-2023 task 6: Using a frustratingly easy domain adaption for a legal named entity recognition system. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1783–1790, Toronto, Canada. Association for Computational Linguistics.

Raphaël Chevrier, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, and Christian Lovis. 2019. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. *Journal of Medical Internet Research*, 21(5):e13484.

Kevin Clark and Christopher D. Manning. 2016a. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016b. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Elliot, Elaine Mackey, and Kieron O'Hara. 2020. *The Anonymisation Decision-Making Framework: European Practitioners' Guide*, 2 edition. UKAN Publication, Manchester.

Diego Garat and Dina Wonsever. 2022. Automatic Curation of Court Documents: Anonymizing Personal Data. *Information*, 13(1).

Christopher Graham. 2012. Anonymisation: managing data protection risk code of practice. Technical report, Information Commissioner's Office, Wilmslow, UK.

Ajay Gupta, Devendra Verma, Sachin Pawar, Sangameshwar Patil, Swapnil Hingmire, Girish K. Palshikar, and Pushpak Bhattacharyya. 2018. Identifying Participant Mentions and Resolving Their Coreferences in Legal Court Judgements. In *Text, Speech, and Dialogue*, pages 153–162, Brno, Czech Republic. Springer International Publishing.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named Entity Recognition in Indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Joffrey L. Leevy, Taghi M. Khoshgoftaar, and Flavio Villanustre. 2020. Survey on RNN and CRF models for de-identification of medical free text. *Journal of Big Data*, 7(1):73.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv.

Xiaoqiang Luo. 2005. On Coreference Resolution Performance Metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Miranda Mourby, Elaine Mackey, Mark Elliot, Heather Gowans, Susan E. Wallace, Jessica Bell, Hannah Smith, Stergios Aidinlis, and Jane Kaye. 2018. Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. *Computer Law & Security Review*, 34(2):222–233.

Arttu Oksanen, Minna Tamper, Jouni Tuominen, Aki Hietanen, and Eero Hyvönen. 2019. ANOPPI: A Pseudonymization Service for Finnish Court Documents. In *Legal Knowledge and Information Systems. JURIX 2019: The Thirty-second Annual Conference*, volume 322 of *Frontiers in Artificial Intelligence and Applications*, pages 251 – 254, Madrid, Spain. IOS PRESS.

Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. 2022. Bootstrapping Text Anonymization Models with Distant Supervision. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4477–4487, Marseille, France. European Language Resources Association.

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.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tom Schamberger. 2021. Customizable Anonymization of German Legal Court Rulings using Domain-specific Named Entity Recognition. Master's thesis, Technical University Munich, Munich, Germany.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. *Journal of Biomedical Informatics*, 75S:S4–S18.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of biomedical informatics*, 58 Suppl(Suppl):S20–S29.

Latanya Sweeney. 2002. K-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, Edmonton, Canada.

Shubham Toshniwal, Allyson Ettinger, Kevin Gimpel, and Karen Livescu. 2020. PeTra: A Sparsely Supervised Memory Model for People Tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5415–5428, Online. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland.